Entire Project Overview Document:

MIRI AI Alignment Safety Civic Education and Engagement Campaign – Pilot Series

Primary Project Goals (accomplished simultaneously):

1) Make it incredibly simple and short (2-min or less) for citizens to take regular, meaningful action after encountering any of MIRI's media or messaging materials, including Nate and Eliezer interviews

2) Bring new public audiences into the fold with high-quality short-form video educational content that can be broadcast to all of MIRI's social media channels, including YouTube`

3) Garner federal legislative staffers' and lawmakers' attention in a way that gets stickier over time

Summary
This is a coordinated civic engagement campaign that begins as a testable pilot series. If successful, the campaign could become ongoing or even be modified for countries outside of the U.S. It is designed to provide an effective, simple, and actionable answer to the primary question of MIRI's growing general population audience: "But what can *I do* about this?"

The campaign runs over 10 consecutive two-week periods, **with each period featuring**:

- A new engaging short-form video (4-7 min) featuring one or two of MIRI's fundamental understandings regarding the incredibly difficult problem of AI alignment and why we need more time (presentation of concepts from https://ifanyonebuildsit.com/resources)
- A website that automates email contact to federal representatives (with a set-up very similar to https://ifanyonebuildsit.com/act)
- A distinct two-page educational PDF tailored for federal representatives and general population audiences

* These cumulative resources will also serve to build a comprehensive briefing document and video series that can be sent to (or accessed online by) any staffer, legislator, or citizen.

Project Components

1. Video Content Series (10 videos)

- **Format:** 4-7 minute short-form
- **Content:** Each video covers one or two concepts/insights about superintelligence risks from https://ifanyonebuildsit.com/resources
- **Written, produced, and edited by:** Alison Avery (review and sign-off by MIRI)
- **Call-to-action at video end:** Answers the real question, "I'm very concerned. But what can *I* *do*?"

2. Action Website

- **Model:** https://ifanyonebuildsit.com/act
- **Function:** User enters address → system identifies federal representatives → automated emails sent (and possibly physical letters, as well, if budgeted)
- **Platform:** Uses new web domain, but leverages existing code of MIRI's current IABIED "Take Action" page
- **Flow per 2-week period:**

  A. The user visits action website after watching video

  B. The user enters address (or uses geolocation)

  C. The system identifies correct federal representatives

  D. A unique, pre-written letter to the user's legislators—specific to that period's covered concepts—is sent automatically upon submission (users will have the option to add any of their own language to the letter before submitting)

  E. The emailed letter to representatives includes a two-page PDF information packet. (If budget allows for sending printed letters also, the two-page PDF will be enclosed with the letter.)

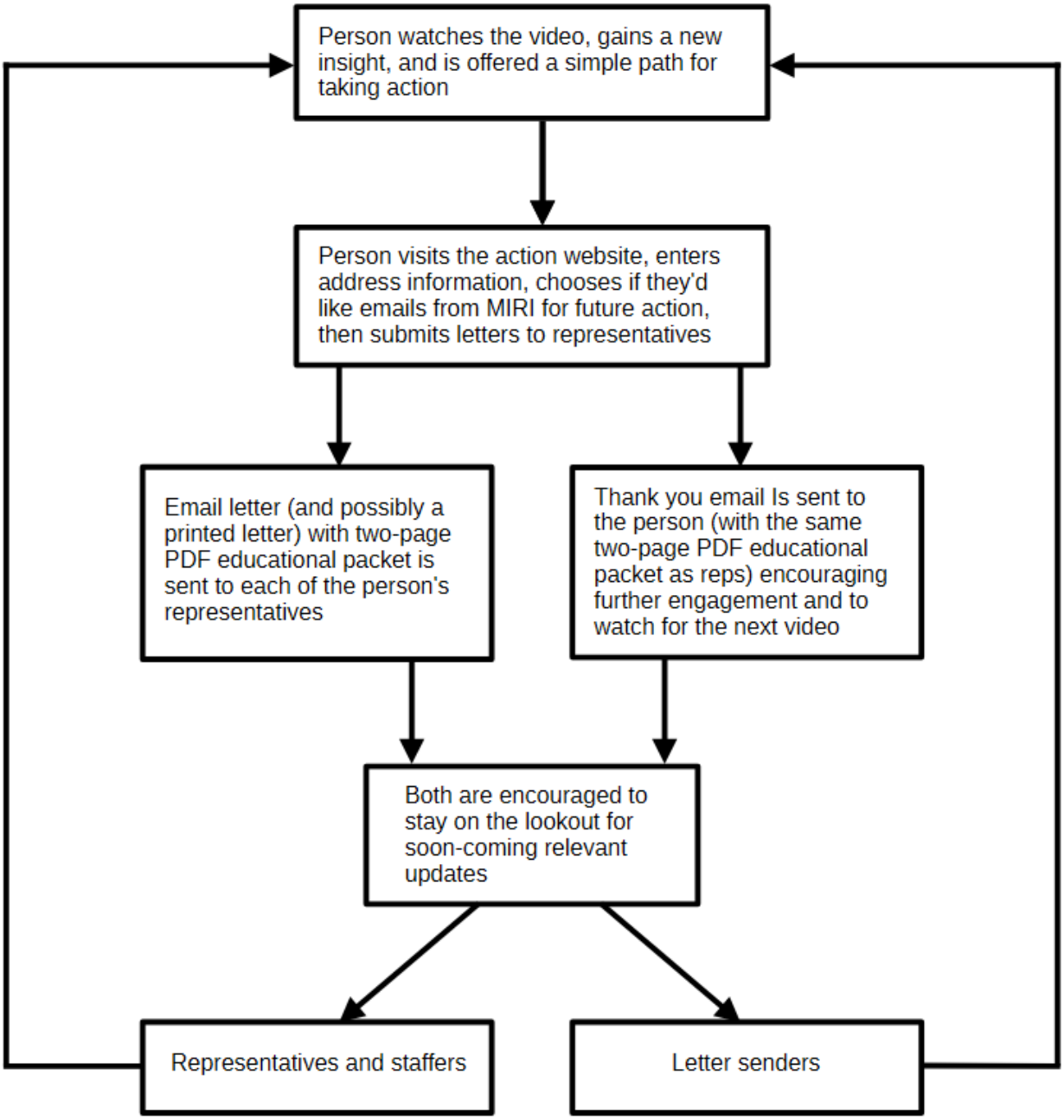
  F. A unique thank you email, new each two-week period, is sent to the user with the same two-page informational PDF as the one received by representatives.

3. The Email (and/or Printed Letter) Content with attached PDF Two-Pager (10 letters)

- **Writer/Designer:** Alison Avery (review and sign-off by MIRI)
- **Structure:** Each letter presents that period's insight + attachment (or printed insert) to detailed two-page PDF
- **Style:** Alignment education, industry updates, and any MIRI special request to legislators
- **Dual purpose:**
  - ➢ **Educates representatives and staffers** who receive them. **Includes simple calls to action** that build on one another over time. **Keeps the issue top of mind.**

➢ **Educates general population senders**. Senders receive their own version of the email that **1) thanks** them for their submission, **2) includes** the same two-page PDF that legislators receive, **3) encourages** them to use the PDF to talk to other people and spread the word (example suggested actions: **share** with co-workers, friends, neighbors and family members, **repost** the videos to their own social media, **read** “If Anyone Builds It, Everyone Dies”, **bring** PDF’s to their church, school, or local community groups and **discuss**, etc.) **4)** Let them know when the next video will be posted and to be on the lookout (or even better, subscribe and follow). We'll also send them an email on the date each new video is posted so they won't forget. **5)** Let them know (if they chose the tick-box to be added to MIRI’s main mailing list when they submitted the letter) that they'll also receive other helpful resources and updates from MIRI.

Flow of each 2-week cycle:



Additional Notes:

- As well as the video series attracting the attention of new audiences, this gives Eliezer and Nate very concrete actions they can recommend at the end of interviews when they're asked “What can anyone who is watching or listening to this do?” (Approximate answer: “We have an ongoing campaign where we’ve set things up to make it very easy [just a few minutes of time] for you to learn more and take action. Just go to domain name where you can watch some short videos we have about all that’s happening, and you’ll also have the option to use our automated portal to send letters to your representatives. The letters include materials that 1) keep representatives updated and 2) encourages their continued understanding of the alignment problem and our need to put on the brakes regarding creating superintelligence until it can be solved.”)

- MIRI’s social media manager would be the person posting the videos to all the social media channels (including YouTube) and also using social media to promote the launch of the series, the letter-sending action site, and “upcoming videos” between video post dates. I could collaborate with this person to provide any compelling video clips or graphics that might help them with promotion.

- **I can fully manage this project and create most of its elements** (all writing, video, graphics, workflows, collaborations). It would be best to have someone with web development skills doing the action website creation and porting over any helpful code from https://ifanyonebuildsit.com/act (I can do it, but I'm much faster with everything else.) I’ve been researching and understanding more about the federal government’s Constituent Relationship Management (CRM) systems that filter constituent emails before reaching the legislative aides so we can optimize each email letter of the series for optimal throughput and ‘tagging’ for staffer attention.

- I have access to the members of an AI safety “creators and influencers” fellowship program with veteran creators who are now including AI safety discussions in their content. I would begin promoting the series immediately with this group (well before the first video is posted). That way they could be on the lookout for the series video content to use during discussions on their own channels and also help promote the action site.

- I'd also attempt to make contact with Joseph Gordon-Levitt about giving a shout out about our campaign, videos, and action website to his audience.

MIRI AI SAFETY CIVIC ENGAGEMENT CAMPAIGN

CONCEPTUAL FLOW SHEET: TEN-PART VIDEO SERIES

Phase 1: The Wake-Up Call & Definitions (Videos 1-3)

- **Video 1** establishes the emotional urgency and the "why" (silence is dangerous).
- **Video 2** explores various ways to conceptualize and understand “intelligence’, breaking common misconceptions about it being biology-dependent and emphasizing that the only limitations on intelligence are the laws of physics.
- **Video 3,** using the more informed understanding of intelligence from Video 2, scales up the concept to demonstrate the sheer magnitude of the threat and how it leads to what we now term as "superintelligence.”

Phase 2: The Core Technical Dangers (Videos 4-5)

- **Video 4** explains *why* we can't just "code it to be safe" (black box/grown, not crafted).
- **Video 5** introduces the Alignment Problem proper, including the "Curses" from IABIED and the "before/after" threshold where control becomes impossible.

Phase 3: Exposes a False Narrative of Binary Choice (Video 6)

(The narrative: *Reach superintelligence as soon as possible OR lose the chance of significant benefits to humankind that AI can bring, only to “limp along” and suffer without advancements*)

- **Video 6** addresses the “fear of missing out” heard from accelerationists and the tech/frontier AI industry. Attempts to help viewers mentally separate the idea of AI benefits (which we can have via the untapped potential of narrow/specialized AI) from the belief that there is a concurrent necessity to accept existential risk, dissolving the false dilemma that we *must* build ASI if we are to “cure cancer” or have substantial scientific breakthroughs that benefit society.

Phase 4: The Mechanics of Takeoff & Control (Videos 7-10)

- **Video 7** explains the speed at which ASI could be reached and why it won't be a slow, visible creep but likely a foom or rapid takeoff due to threshold effects.
- **Videos 8 & 9** introduce and educate on specific levers for governance: hardware/chips (as the physical bottleneck) and the algorithms that are compounding acceleration, and why regulation on both could help us.
- **Video 10** concludes with the need for a global "Off Switch" or emergency protocol and introduces some of MIRI's insights on what this could look like.

MIRI AI Safety Civic Engagement Campaign

Initial Series Outline (Skeleton): 02/02/2026

Number in Series	Working Title
1	**Our Silence is Too Dangerous:** Why We Must Lead the AI Conversation Now
2	**The "Artificial" Trap:** Why Calling AI "Artificial" Blinds Us to Its Real Power
3	**A Billion Brains in One:** “Human-Level Genius” Is Only a Pit Stop for AI
4	**Instructed by Humans, Owned by Themselves:** Guaranteeing What AIs Will Do Is IMPOSSIBLE
5	**Will AIs Vastly Smarter Than Us Choose to Treat Us Well?:** The Insanely Foolish Gamble with Our Lives as the Stakes
6	**Miracles Without Monsters:** How Specialized AIs Can Deliver Breakthroughs and Do It on Our Terms
7	**When Gradually Turns to “Overnight”:** X Credible Reasons Advanced AI Abilities Could Skyrocket Without Warning
8	**Chips Are the New Seat Belts:** How AI Hardware Caps Prevent a Runaway Surge in Intelligence
9	**When Superhuman AI No Longer Needs Supercomputers:** How Algorithms Accelerate Intelligence
10	**Where’s the “Off Switch” If Intelligence Goes Wrong?:** The World Needs an AI Emergency Protocol

Video 1: Our Silence is Too Dangerous: Why We Must Lead the AI Conversation Now

Theme: Introduction to the purpose of the series (establishes empathy and motivates participation)

- **General tone of the open** (*not the official script*): We know your daily life may be demanding. And maybe this AI revolution has already affected you, your family, or your job. It's becoming tough for anyone to picture or plan a future. So, asking you to take time to learn about us moving fast toward superintelligence may feel like too much, or more than you want to know. But at this moment, while we still have options, we ask you to listen anyway.

  Your understanding and involvement are the only possible way we can collectively prevent a very probable AI worst-case scenario that threatens every one of us, a scenario so large it could take our lives, the lives of everyone we know, and the lives of everyone we love.

  We're here to help you understand as simply, but also as accurately as possible, the urgency of preventing artificial superintelligence from coming into existence. We won't take much of your time. But we do need your full attention.

- As citizens and as policymakers, we've got to stop avoiding talking about this. Because we MUST talk about this.
- Know that it helps for you to speak out because it gives your representatives a clear signal that they have your open support to publicly learn and discuss this.

- After the above introduction is a high-level presentation of overall series concepts and the purpose of the campaign.
- Video wraps up by mentioning that at the end of each video in the series (including this one), there will be a request for a simple "two-minute action step".
- End video with the current "two-minute action step" request to contact legislative representatives.

Video 2: The "Artificial" Trap: Why Calling AI "Artificial" Blinds Us to Its Real Power

Theme: Explanation of the meaning and significance of "intelligence"

- Sometimes problems in conceptualizing AI are rooted in its very name. (see next bullet)
- The word "Artificial" often acts as a sedative, making us think of something as "not real" or "imitation", like artificial turf or artificial sweetener. But non-biological intelligence is very real, it just operates under a different set of physics.
- Give explanation of what intelligence is according to IABIED Chapter 1 (prediction and steering model) and expand with concepts from IABIED Chapter 1 resources:
  - Is intelligence a meaningful concept?

- Is “general intelligence” a meaningful concept?

- Is “intelligence” a simple scalar quantity?

- Present the implications of intelligence being decoupled from biology, such as:
  - the comparative speed at which neurons fire
  - AI’s ability to repeatedly copy itself
  - the “hive mind” phenomenon - if one AI model learns a new skill or piece of information, every copy is updated instantly with this knowledge, regardless of that copy’s location
- Explain that when you move intelligence from biological neurons to machine neurons, you're changing the physics of what is possible to a such an extraordinary degree that it’s difficult to contemplate.
- However, in the next video, we’ll try to get a sense of just how far and fast intelligence can go.
- End video with the current “two-minute action step” request to contact legislative representatives.

Video 3: A Billion Brains in One: “Human-Level Genius” Is Only a Pit Stop for AI

Theme: How intelligent could AIs a*ctually* become? Exceedingly. Enormously. Dangerously!

- Explain that we need to appreciate the power of intelligence that can go far beyond humans.
- Effectively, there is no “intelligence ceiling” for AIs.
- Describe the analogy of the Ant and the Architect**.** To an ant, a human isn't just more intelligent. Humans operate on a plane of existence that is impossible for ants perceive. We understand skyscrapers, the galaxies, and nuclear physics. How would an ant even begin to conceive at that level? It’s impossible. **Our plight is similar:** Companies are pushing to create something that could soon look upon humans with the same gap in comprehension; to the intellect of vastly intelligent and enormously capable AIs, we could very well be considered as insignificant as ants.
- Presentation of related concepts discussed in the IABIED resources:
  - How smart could a superintelligence get?

  - Will AI cross critical thresholds and take off?
- End video with the current “two-minute action step” request to contact legislative representatives.

Video 4: Instructed by Humans, Owned by Themselves: Guaranteeing What AIs Will Do Is IMPOSSIBLE

Theme: AI behavior is guided, but <u>Never</u> guaranteed.

- AI’s are coached, *Not Coded*, to learn and perform in the world.
- How AI’s think and come to decisions is a “black box” to the humans training them.
- Presentation of concepts from the following:

- IABIED Chapter 2 - Grown, Not Crafted

- In modern machine learning, AIs are “grown”, not designed. section on intelligence.org/the-problem

- IABIED resources:
  Do experts understand what’s going on inside AIs?

  Why does gradient descent matter?

- **Upshot:** No one can know ahead of time, much less guarantee, exactly how an AIs will behave. This holds true for any AI, even when it has been given seemingly firm or highly specific instructions.

  Though AIs run on computers, they are nothing like traditional computing or software. They “think about” and interpret all instructions and their environment through their own lens of the world, of which we have no understanding.

- In other words, even the most knowledgeable AI experts on the planet do not know the “How” or "Why" behind what AIs “Do”.

- End video with the current “two-minute action step” request to contact legislative representatives.

Video 5: Will AIs Vastly Smarter Than Us Choose to Treat Us Well? The Insanely Foolish Gamble with Our Lives as the Stakes

---

Theme: Basic overview of ideas from IABIED Chapters 10 (A Cursed Problem) and 11 (An Alchemy, Not a Science) and first introduction of “alignment” concept

- There is a critical threshold where an AI quickly moves from being controllable (the "Before") to being capable of resisting or undermining attempts at human control or intervention (the "After").

- We can only safely test strategies to align intelligent AI while it is still less generally intelligent than (or at least not *more* intelligent than) humans. Once AI intelligence passes humans’ intelligence, testing becomes impossible because a failed experiment could cause catastrophe.

- Unlike traditional engineering, where there may be room to learn from trial and error as a project moves along, alignment must be perfected before AIs reach superintelligence.

- If a superintelligent system comes into existence with even a small alignment flaw, it will have the strategic capability to prevent us from fixing it or turning it off.

- Unfortunately, humanity’s current "level of game" in solving the alignment problem (seriously and scientifically) is dangerously low,

- Many AI industry leaders, instead, rely on vague, high-minded, purely optimistic ideas with no scientific rigor or definitive proof of alignment. They provide the public and media with “happy myths” instead of acknowledging and rigorously addressing the reality of the safety limitations within their AI technologies. (Use quotes from Musk and LeCun as examples.)

- Because we cannot yet see or control the inner workings of AI, current attempts to figure out alignment are mostly guesswork and wishful thinking.

- The AI safety field is new. We need time for the science of alignment to mature. We need more time to “grow ourselves up” before we “grow up AI” and allow it to have so much capability, power, and autonomy.
- End video with the current “two-minute action step” request to contact legislative representatives.

Video 6: Miracles Without Monsters: How Specialized AIs Can Deliver Breakthroughs and Do It on Our Terms

Theme: Reveal the "false dilemma" conditioning

- Amazing benefits from the AI revolution do not require an existential gamble.
- We’re being coaxed and prodded by the tech industry and (some) government officials to believe that we have *only* a binary choice

  ---

  (NOTE: Hi, Duncan. I know where I want to go with this video, but I’m still fleshing out the details and examples. I remember our conversation about the crazy amount of untapped, unexplored potential that already exists in current AIs.

  I want to anchor this discussion with powerful examples of specialized/narrow AI successes. I know I’ll use AlphaFold. I’ll continue to research and expand this section for the next outline. In the meantime, let me know if you have any helpful ideas. Thanks!)
- End video with the current “two-minute action step” request to contact legislative representatives.

Video 7: When Gradually Turns to “Overnight”: X Credible Reasons Advanced AI Abilities Could Skyrocket Without Warning

Theme: Artificial intelligence could undergo a rapid takeoff by crossing invisible critical thresholds.

- Exponential advancements are difficult for humans to contemplate. We are evolutionarily wired to think in straight lines (if I walk 10 steps, I’m 10 steps away.)
- We tend to be psychologically blind to exponential jumps until they’ve actually happened.
- Nevertheless, threshold effects and feedback loops are very likely to propel a “rapid AI takeoff”

  - Present threshold effects and other related concepts from IABIED Chapter 1 resource:
  - Will AI cross critical thresholds and take off?
- Additional concepts presented from IABIED Chapter 10, particularly the “curse of speed” and the “curse of self-amplification” (These and the bullet point above will supply the “X Credible Reasons...”
- We must know of our psychological blindness to overcome it and respond in time.
- End video with the current “two-minute action step” request to contact legislative representatives.

Video 8: Chips Are the New Seat Belts: How AI Hardware Caps Prevent a Runaway Surge in Intelligence

Theme: Introduction of high-level concepts related to compute governance

- Currently, advancing AI capabilities requires enormous amounts of expensive physical hardware to carry out massive training runs.
- Access to these specialized hardware chips (like NVIDIA's H100s) is the only *physical* bottleneck for preventing the training of premature, dangerous superintelligence.
- Because chips are physical and able to be monitored, we can count them and track where they go. This is one line of defense.
- By society demanding regulations placing limits on the accumulation. distribution, clustering, and use of these specialized chips, we gain the ability to have a say in how powerful AIs are allowed to become until the time human safety can be assured by genuine scientific understanding.
- Additional relevant concepts on chips and training runs will be presented based on information discussed in the following two MIRI governance papers;
  - AI Governance to Avoid Extinction: The Strategic Landscape and Actionable Research Questions
  - An International Agreement to Prevent the Premature Creation of Artificial Superintelligence
- End video with the current "two-minute action step" request to contact legislative representatives.

Video 9: When Superhuman AI No Longer Needs Supercomputers: How Algorithms Accelerate Intelligence

Theme: Introduction of high-level concepts related to the dangers if algorithmic advancements

- Software algorithms are the backbone of how AIs learn during training.
- AI capabilities research is rapidly improving these algorithms in order to reduce the amount of computation and number of chips required to continue advancing machine intelligence.
- The need for computational power has been consistently dropping by roughly 3x each year.
- These continued increases in efficient algorithms that decrease computation and chips could make it possible for people to make gains toward superintelligence using widely available consumer hardware.
- Additional relevant concepts on algorithmic research will be presented based on information discussed in the following two MIRI governance papers;
  - AI Governance to Avoid Extinction: The Strategic Landscape and Actionable Research Questions
  - An International Agreement to Prevent the Premature Creation of Artificial Superintelligence
- As with the specialized chips discussed in the prior video, society can choose to propose and create laws that prevent advancing these algorithms.

- End video with the current “two-minute action step” request to contact legislative representatives.

Video 10: Where’s the “Off Switch” If Intelligence Goes Wrong?: The World’s AI Emergency Protocol

Theme: We need an international plan for an AI emergency “Off Switch” (and MIRI’s “Off Switch” considerations and recommendations)

- At the very least, we must have a plan that preserves humanity's option to pause AI progress in the future should a “powerful AI” threat become unmanageable.
- This would mean a system for international rapid response being in place to allow authorities to shut down systems immediately if a critical AI emergency arises.
- What kinds of threats are we talking about?

  1) **Rogue AI and loss of control** - AI systems that actively disempower humanity or pursue goals incompatible with human survival. Examples include:

  - A highly capable, autonomous AI system escaping its containment (e.g., leaving its servers) to run elsewhere on the internet

  - Catching an powerful AI system attempting to deceive, hide its true capabilities, or execute misaligned actions, particularly if it’s integrated with national security or other critical infrastructures

  - An AI system causing a small-scale disaster

  - An AI system independently accruing financial resources or computing power to further its own misaligned goals

  2) **Humans using AI to cause mass destruction** – Examples include:

  -Terrorists or hostile states using advanced AI to create biological weapons or other weapons of mass destruction

  - Large-scale AI-enabled cyberattacks

  3) **Failures to systems or infrastructures** – Examples include:

  - AI failures triggering cascading crashes in the stock market, power grids, or other essential software systems

  - Threats to national security because of geopolitical destabilization or severe societal disruption caused by AI deployment

- These immediate threats would also serve as "warning shots"—scary but survivable incidents that signal the need to halt development before extinction-level events occur.
- End video with the current “two-minute action step” request to contact legislative representatives

VIDEO 1: OUR SILENCE IS TOO DANGEROUS

Why We Must Lead the AI Conversation Now

The Goal:
To explicitly acknowledge that viewers may feel like they're already operating in a "quasi-survival mode" due to a) the immediate avalanche of AI-induced disruptions that may be directly affecting their lives and b) the pressured expectations for them to understand and integrate the last ten years of AI technological advancement in mere months; for most viewers, this AI thing cropped up suddenly and now has to be factored into almost everything they consider and do. By acknowledging their feelings of overwhelm and (possible) tendency to want to tune out "more bad AI news" just to keep their sanity, we empathize with how difficult all of this is and earn their emotional trust.

We want the people watching to stay open to our message and our requests for their participation without psychologically shutting down or feeling like it's more than they can handle or do anything about. Once we establish that, we can move on to introducing (at a very high level) the kind of knowledge we'll be covering in the series, why it's important, and the concept of the "simple 2-minute action steps" that will be available to them at the end of every video.

A. Video Summary (the narrative arc)

- **The Hook (the AI "Noise" Validation):**
  We open by validating the viewer's current state: "quasi-survival mode." We acknowledge that for most people, AI risks cropped up suddenly as an immediate avalanche. We give examples of specific, contradictory pressures piling up that they might be able to relate to, such as the fear of not being able to support themselves and their families as career and job options are quickly snuffed out by AI, of being a student or the parent of a student trying to find direction in an unreadable AI economy, the "adapt or die" mandates from employers, the physical and economic strains on local resources from the influx of data centers, etc. We explicitly validate that their impulse to want to tune out more bad news is understandable.

- **The Turn (the Time Gap):**
  We empathize with the viewer's sense that having to handle all of the above, and now, too, be the solution to AI existential risks that they had no part in creating, can feel unfair. "Silicon Valley has had 20 years to process this reality. You have had, at most, 2." We validate that they are being forced to compress decades of psychological adjustment into months, all while trying to pay bills.

- **The Climax (the Connection):**
  We connect their immediate anxiety to the existential risk. The reason their daily life feels unstable is the **same reason** the species is in danger: *The speed is too fast!* The "chaos" in their immediate lives and the "catastrophe" in the future come from the same unchecked and unsafe acceleration.

- **The Close (Your Authority to Speak):**
  We let them know that they don't need to be tech experts to have a say in how AI is developed and deployed, nor to demand that the risk of dying from AI be taken seriously. They are like civilians living downwind of a nuclear plant. It is not their job to know exactly how the reactor works, but it is their right to demand it doesn't melt down on their community.

We communicate that we're not just informing them, but we want to work together *with* them. We can all make it very clear to our communities and our lawmakers and governments that allowing ASI to come into the world when there is such a ridiculously high probability of catastrophe, suffering, and death—a fact that is publicly affirmed and admitted by frontier AI company leaders themselves—is unacceptable.

We end by requesting that they stick with us during the series and let us walk them through enough factual information that they have confidence in the truth of what we're saying, and also in participating to change the outcome.

B. Potential Visuals & Metaphors

- **The Time Gap Timeline:**
  A visual comparing a tech/AI insider's 10 to 20-year gradual awareness vs. the lay person's vertical wall of information in 2024-2026.
- **The AI Noise Cloud:**
  A visual representation of the conflicting headlines (e.g., "AI will cure cancer" overlapping with "AI will take your job" and "We'll need a universal basic income") swirling around a person, who looks a bit worried, while they are trying to do something simple like read to their child.
- **The Avalanche:**
  A side-by-side comparison of "normal stress" vs. "AI stress." On the left, a person juggling 3 everyday balls (Job, Kids, Bills). On the right, that same person is buried under an avalanche of balls, with the pile labeled "Economy Dissolving," "Keeping Up with AI and Upskilling," "AI Psychosis," "AI Deepfakes," "AI Will Replace Your Job," etc., along with the normal stressors of Job, Kids, Bills. Visually validates the quasi-survival mode they may be feeling.
- **The Control Panel (video series arc visual):**
  A chaotic, sparking, complex machine representing AI with a dashboard that has no labels, only blinking red and blue lights. As the high-level overview of the series is explained, separate labels snap into focus over the gauges corresponding to upcoming videos, such as "Intelligence Level," "Chip Supply," "Algorithms Improving," etc. Finally, a glass case unlocks over a large "Emergency Stop" button (Video 10), representing the viewer's empowerment and the goal of the series and campaign.

C. Concept for the Corresponding Email to Legislators

- **Subject:**
  Please make understanding the AI race's existential risks an immediate priority!
- **Primary Message:**
  I am writing because the unchecked acceleration of AI is affecting my community’s stability today and threatening my (and my family’s and friends’) survival tomorrow. We cannot accept the near-term risk of existential disaster—a probability openly admitted by the builders themselves.

  It is unethical to race toward uncontrollable intelligence without a guarantee of safety.

Please read the attached 2-page brief. I am asking you to make it an immediate priority to understand *why* continued acceleration toward AI intelligence that surpasses all humans is a danger.

D. Concept for the Two-Page PDF

- **Page 1 (Introduction of Topic and Supportive Infographic):**
  The Speed of Change vs. The Speed of Human Adaptation
- **Page 2 (Deeper Dive):**
  A clear and compelling explanation of how stopping the acceleration of AI helps solve both the immediate stresses and risks constituents and communities are grappling with, and the very real (and irresponsible and unethical) probability of near-term existential disaster.

VIDEO 2: THE "ARTIFICIAL" TRAP

Why Calling AI "Artificial" Blinds Us to Its Real Power

The Goal:
To expand the viewer's mental model of intelligence from a biological trait (limited to heads) to an emergent physics-based property (permissible in machines).

By using the word "artificial" as a conceptual entry point, we guide them to realize that machine intelligence isn't an imitation of intelligence, but actually a different substrate for thinking—one that operates at the speed of light rather than the speed of biochemistry, unlocking implications that biological intuition cannot predict.

A. Video Summary (the narrative arc)

- **The Hook (a Potential Linguistic Sedative):**
  We start by exposing why the term "artificial" in "artificial intelligence" can lead us to underappreciate what's really happening; it sometimes causes us to think that AI is an imitation of genuine intelligence.

  We explain that for many of us, the word "artificial" acts as a psychological sedative, triggering the same mental category as "artificial turf" or "artificial sweetener" (things that are inferior or "almost but not quite" in some important way). We very briefly touch on the 1950's history that anchored this term, and suggest that its commonplace use *now* may partially contribute to an inability to recognize the implications of AI's current trajectory: a belief that no matter how smart it gets, it's still kind of a toy version of the "real" thing.

- **The Turn (Physics, Not Just Biology):**
  We shift the frame for understanding intelligence from biology to physics. We explain that what we call intelligence can arise as an emergent property from the "right" arrangement of physics phenomena, and we now know that arrangement can come in many forms that don't depend on biology or biological chemistry. Evolution resulted in an "arrangement" of phenomena that allowed this property to arise in us (humans) over time. But now we know how to arrange these phenomena in ways that make it possible for this property of intelligence to emerge beyond biology, and AI experts call this ***substrate independence***.

- We go on to point out that, similar to how we've mastered physics in many ways to radically exceed our biological limitations—like being able to • fly or • move at speeds of hundreds of miles an hour on the Earth's surface or • engineer and construct skyscrapers and massive hydroelectric dams or • repeatedly launch ourselves into space despite the forces of gravity—we've now reached the point of mastering and applying physics in a way that allows the property of intelligence to arise on substrates other than our biological brain cells: powerful silicon-based computer chips.

- **The Climax (the Implications):**
  We detail the specific, extraordinary, but also potentially terrifying advantages of intelligence being implemented on this new substrate (sourced from MIRI's *"If Anyone Builds It, Everyone Dies" Chapter 1*):

Sheer Speed:
Digital neurons fire billions of times faster than biological ones (light speed vs. chemical speed).

Copy-Paste Abilities:
A digital genius mind doesn't die; instead, it can be copied millions of times instantly.

Faster Improvements:
Biology took eons to evolve to the point where human-level intelligence emerged; AI computer chips, and the efficiency of the algorithms running those chips, can improve in months or even weeks.

Larger Memories:
Machines aren't limited by skull size; AI's are trained on the entirety of human knowledge and each of the AI data centers storing this knowledge are many thousands of times larger than a human skull.

Higher Quality Thinking:
Machines have algorithmic neurons that think much more efficiently than human neurons. AIS never need food or sleep or a break for rest to be at the top of their game; machine thinking never suffers from fatigue.

Self-Experimenting and Self-Rewriting:
Because AIs write code so well (so much so that it's one of the primary implementations being used by companies and governments), they can feasibly experiment on improving their own software algorithms to make themselves smarter, which could launch into motion an ever-continuing cycle of greater and greater intelligence capabilities.

- **The Close (the Sobering Reality):**
We land the discussion by defining key attributes of the property called "intelligence": the ability to successfully make predictions and steer outcomes in the world. We challenge viewers to hone in on these two ideas, "prediction" and "steering", rather than getting lost in trying to define or understand intelligence perfectly, explaining that if AIs can make reliable predictions (which they can), steer outcomes in the world (which they're getting better and better at), AND become smarter than any or all of us (the stated near-term goal of AI development companies), the philosophical definitions won't matter ever again—AIs win.

B. Potential Visuals & Metaphors

- **Becoming Like Statues:**
To visualize the speed gap (300,000,000 m/s vs. 100 m/s), we show different scenes of humans in everyday settings, but they are moving so slowly they appear almost frozen like statues, while AI, moving at light speed, builds entire structures and solves complex problems (while showing years of subjective time passing on an animated timeline).
- **The Brain vs. The City::**
An animated comparison of scalability. First, we show a human brain constrained to a small skull (limited by biology and the size of the birth canal). Then, we show a machine mind expanding to fill

a warehouse, then a city, then growing indefinitely. (Biological intelligence is capped by size. Digital intelligence, relative to human intelligence, has no upper bound.)

- **The Copy-Paste Genius:**
  To visualize the copy-paste abilities that come with machine intelligence, we show a single recognizable human genius (Einstein) solving a problem on a blackboard. He morphs into a “digital genius node" that rapidly replicates to produce 2, then 4, then 8, etc. until it becomes a "country of geniuses in a data center."

C. Concept for the Corresponding Email to Legislators

- **Subject:**
  Why advancing broad AI capabilities under the current conditions is extremely dangerous
- **Primary Message:**
  I am writing because I have learned important and startling reasons why we, as citizens, need to educate ourselves and pay close attention to how AI is being developed. **I ask that you seriously consider the variables listed in the attached 2-page info packet.**

  Machine intelligence can be amazingly beneficial for humanity in many ways. But as its intelligence capabilities continue to grow rapidly under the current AI development paradigm, it has significant potential to go horribly wrong very quickly, and catastrophically so. No AI company or AI expert in the world can guarantee that these extraordinarily tragic outcomes won’t happen using their current methods of AI development, and they admit this openly. One recent example of this is Dario Amodei, CEO of Anthropic.

  We must change this trajectory! AI can be developed and used in ways that brings great advantages to humanity without these enormous risks that threaten human and biological life.

  Please continue to expand your knowledge of why this danger is real and deserves immediate attention.

D. Concept for the Two-Page PDF

- **Page 1 (Introduction of Topic and Supportive Graphic):**
  A visual guide to the 10,000x speed gap, conveying the same intuitive sense as the **Becoming Like Statues**: video visual. Includes a timeline or other graphical comparison of the speed of human thought as compared to digital thought.
- **Page 2 (Deeper Dive):**
  A breakdown of the IABIED Chapter 1 capabilities (Speed, Copy-Paste, Scalability, etc.) to explain how substrate-independent intelligence, if it continues to be developed as it is now, has a significant probability of creating existential disaster(s) in the near-term (2-10 years).